Identification of "Missing Link" Families of Small DNA Tumor Viruses

Nicole L. Welch[1], Michael J. Tisza[1], Anna Belford[1], Diana V. Pastrana[1], Yuk-Ying S. Pang[1], John T. Schiller[1], Ping An[2], Paul G. Cantalupo[2], James M. Pipas[2], Samantha Koda[3], Kuttichantran Subramaniam[3], Thomas B. Waltzek[3], Chao Bian[4], Qiong Shi[4], Zhiqiang Ruan[4], Terry Fei Fan Ng[5], Gabriel J. Starrett[1], and Christopher B. Buck[1]*

[1]Lab of Cellular Oncology, NCI, NIH, Bethesda, MD, 20892 USA

[2]Dept. of Biological Sciences, Univ. of Pittsburgh, Pittsburgh, PA 15260 USA

[3]Dept. of Infectious Disease and Pathology, Univ. of Florida, Gainesville, FL, 32611 USA

[4]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, BGI, Shenzhen, Guangdong 518083, China

[5]College of Veterinary Medicine, University of Georgia, Athens, GA 30602 USA

*Corresponding author: buckc@mail.nih.gov

Abstract

Adenoviruses, papillomaviruses, parvoviruses and polyomaviruses are collectively known as small DNA tumor viruses. Although it has long been recognized that small DNA tumor virus oncoproteins and capsid proteins show a variety of structural and functional similarities, it is unclear whether these similarities reflect descent from a common ancestor, convergent evolution, horizontal gene transfer among virus lineages, or repeated acquisition of genes from host cells. We report the discovery of a dozen new members of an emerging fourth family of small DNA tumor viruses, the *Adomaviridae*. Cell culture-based expression of adomavirus open reading frames showed that the virion structural proteins of adomaviruses are homologous to those of adenoviruses. The search for adomaviruses unexpectedly revealed the existence of a previously unrecognized family of animal DNA viruses that are distantly similar to virophages and other polinton-like viruses of unicellular eukaryotes. Members of the emerging "adintovirus" family encode retrovirus-like integrase proteins and adenovirus-like DNA polymerase and virion structural proteins. Adintovirus sequences were found in genomics datasets for all major groups of animals, including terrestrial vertebrates. Analysis of the gene content of adintoviruses supports a model in which small DNA tumor viruses descended from more complex adintovirus-like ancestors through gene-loss and horizontal gene transfer events occurring hundreds of millions of years ago.

Author Summary

In contrast to cellular organisms, viruses do not encode any universally conserved genes. Even within a given family of viruses, the primary amino acid sequences of homologous genes can rapidly diverge to the point of unrecognizability. Although modern metagenomic surveys have begun to expose the wide sequence diversity within known virus families, it is unclear how much of the unidentifiable sequence "dark matter" typically observed in these surveys represents unknown virus groups. Our searches of sequence databases revealed more than a dozen highly divergent new members of an emerging virus family, the *Adomaviridae*. Using structure-guided methods for detection of distant homologs, we also discovered the existence of an apparently older lineage of viruses that we name "adintoviruses." Adintovirus sequences can be found as integrated into host animal germline DNA and as apparently linear DNA molecules in genome sequencing datasets for all major animal phyla. The results suggest a framework that ties small DNA tumor viruses into a shared evolutionary history in which key family-level divergence events occurred in early animals and their unicellular precursors.

Introduction

Polyomaviruses, papillomaviruses, parvoviruses, and adenoviruses are collectively referred to as small DNA tumor viruses. The four virus families encode non-enveloped virion proteins with similar β-jellyroll core folds and express functionally similar oncogenes that inactivate cellular tumor suppressor proteins (Figure 1)[1,2]. It has remained unclear how small DNA tumor viruses might be interrelated. A current evolutionary model has placed adenoviruses on a spectrum of double-stranded DNA (dsDNA) polinton-like viruses that encode hallmark type B DNA polymerase (PolB) and integrase genes. Polyomaviruses, papillomaviruses, and parvoviruses are proposed to have descended from unrelated circular Rep-encoding single-stranded DNA (CRESS) virus ancestors [3]. While this model accurately represents many of the features of modern small DNA tumor viruses, it does not explain a number of observed similarities between the four families. Achieving a better understanding of the evolutionary origins of small DNA tumor viruses has the potential to direct comparative studies of these ubiquitous human pathogens.

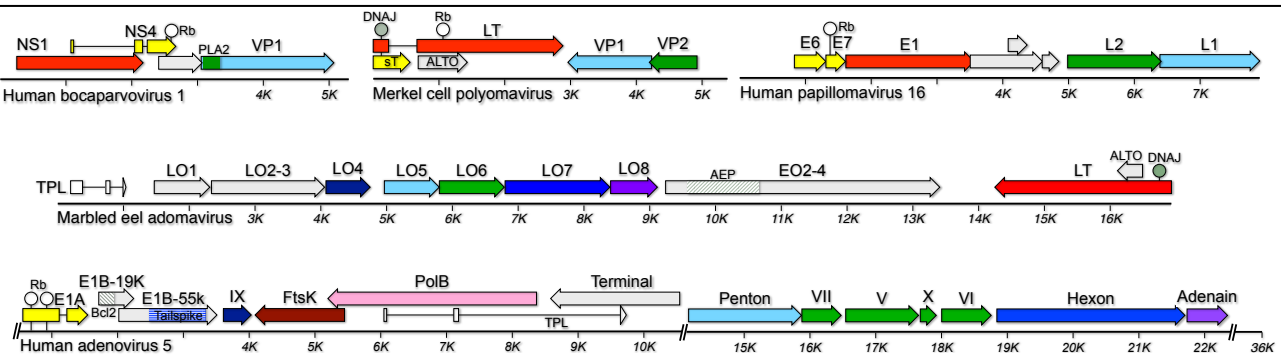


Figure 1: Examples of small DNA tumor virus genomes. Genes are color-coded based on known or inferred functions. AEP: domain resembling archaeal-eukaryotic primase small catalytic subunit; white lollipop: retinoblastoma (Rb) interaction motif; gray lollipop: domain with predicted fold similar to DNAJ chaperone proteins; PLA2: domain with predicted fold similar to phospholipase A2; Tailspike: domain with predicted structural similarity to dsDNA virus tailspike proteins or carbohydrate degrading enzymes. See main text for explanation of other gene names. Supplemental Table 1 lists accession numbers and host animal taxonomic designations.

In 2011, a novel circular dsDNA virus was discovered in a lethal disease outbreak among Japanese eels (*Anguilla japonica*)[4,5]. Two related viruses have since been isolated from Taiwanese marbled eels (*Anguilla marmorata*) and giant guitarfish (*Rhynchobatus djiddensis*)[6,7]. The name "adomaviruses" has been applied to this emerging family based on the fact that they encode proteins with distant similarity to the virion-maturational proteases of *ad*enoviruses as well as proteins distantly similar to the replicative superfamily 3 DNA helicases (S3H) of poly*oma*viruses, papill*oma*viruses, and parvoviruses [8].

The starting goal of the present study was to identify the adomavirus major and minor capsid proteins. To discover additional adomaviruses, particularly those that might serve as laboratory models for adomavirus infection, we performed metagenomic sequencing and developed a bioinformatics pipeline to detect candidate adomavirus sequences in the NCBI's Sequence Read Archive (SRA). In addition to identifying more than a dozen adomaviruses, the searches serendipitously revealed the existence of a previously unrecognized virus family that encodes

82 distinctive PolB and virion proteins that are distantly similar to those of adenoviruses and
83 parvoviruses. Members of the previously unrecognized family also encode a retrovirus-like
84 integrase gene related to the integrase genes of a group of unicellular eukaryote-associated
85 polinton-like viruses known as virophages [9]. Analysis of the gene content of the emerging
86 "adintovirus" family sheds light on the evolutionary history of small DNA tumor viruses.

87 Results

88 Identification of adomavirus virion proteins

89 In a gene-naming convention developed for Japanese eel adomavirus (NC_015123), hypothetical
90 proteins are numbered using inferred late-strand open reading frame (LO) and early-strand open
91 reading frame (EO) designations. HHPred searches revealed that adomavirus LO8 proteins
92 resemble the virion maturational proteases of adenoviruses. This suggests the hypothesis that
93 adomavirus LO4-7 genes might be syntenic homologs of adenovirus virion structural proteins.
94 We set out to experimentally test this hypothesis.

95 To determine whether predicted adomavirus LO proteins are expressed from spliced or unspliced
96 ORFs, RNAseq data published by Wen and colleagues [6] were analyzed to determine the splicing
97 patterns of marbled eel adomavirus transcripts. Splice acceptors immediately upstream of the
98 inferred ATG initiator codons of LO4, LO5, LO6, and LO7 proteins were extensively utilized
99 (Supplemental Table 2). Messenger RNAs encoding the adenovirus late genes carry a tripartite
100 leader (TPL) that has been shown to enhance translation [10]. A similar three-exon leader sequence
101 was detected upstream of the marbled eel adomavirus LO genes (Figure 1).

102 HHpred searches were performed to detect predicted structural similarities between adomavirus
103 and adenovirus proteins. LO6 ORFs show high-probability (>95%) predicted structural similarity
104 to a 37 amino acid hydrophobic segment of adenovirus pX, a minor virion core protein that is
105 thought to participate in condensation of the viral chromatin [11]. LO6 alignments also showed
106 moderate probability hits for adenovirus pVI, which is believed to play a role in destabilizing
107 cellular membranes during the infectious entry process [12]. The results suggest that LO6 is a fused
108 homolog of adenovirus pVI and pX virion core proteins.

109 Adomavirus LO4 proteins have a predicted C-terminal trimeric coiled-coil domain. This
110 relatively generic predicted fold gives a large and diverse range of hits in HHpred searches,
111 including the coiled-coil domains of various viral fiber proteins (for example, avian reovirus
112 σC). Negative-stain EM findings indicate that adomavirus virions do not have a discernible
113 vertex fiber [4,6,7]. Some individual LO4 proteins, as well as alignments of LO4 proteins, yielded
114 HHpred hits for adenovirus pIX, a trimeric coiled coil protein that serves as a "cement" that
115 smooths the triangular facets of the adenovirus virion. The hypothesis that LO4 is a pIX homolog
116 could explain the smooth appearance of adomavirus facets in negative-stain EM.

117 Although LO5 and LO7 did not yield informative results in BLAST, HHpred, or Phyre[2]
118 searches, the inferred homology of the adomavirus LO4, LO6, and LO8 genes to minor
119 components of adenovirus virions led us to hypothesize that LO5 is a syntenic homolog of
120 adenovirus penton (a pentameric single-jellyroll protein that occupies virion vertices) and that
121 LO7 is a syntenic homolog of adenovirus hexon (a tandem double-jellyroll protein that
122 trimerizes to form virion facets) [1,11]. To experimentally test this prediction, marbled eel

123 adomavirus was grown in eel kidney cell culture [6] and virions were purified using Optiprep
124 gradient ultracentrifugation. Virion-rich gradient fractions were separated on SDS-PAGE gels
125 and bands were subjected to mass spectrometric analysis (Supplemental Figure 1). The results
126 indicate the presence of substantial amounts of actin and myosin derived from host eel cells.
127 Adomavirus LO2-3 proteins show high probability HHpred hits for Wiscott-Aldrich syndrome
128 protein and other actin-regulatory proteins, suggesting the hypothesis that the virion preparations
129 might be contaminated with host cell actin due to interactions with LO2-3. Only a small number
130 of individual peptides representing LO2-3 were detected by mass spectrometry. The analysis
131 identified prominent bands in the Coomassie-stained gel as LO4, LO5, LO7, and LO8. The
132 relative intensities of bands identified as LO5 and LO7 are consistent with the prediction that the
133 two proteins constitute penton and hexon subunits, respectively. Lower molecular weight bands
134 showed hits for LO6, consistent with the prediction that this protein is present in virions in an
135 LO8-cleaved form. Marbled eel and guitarfish adomavirus LO6 proteins encode potential
136 adenain cleavage motifs ((MIL)XGXG or L(LR)GG) [13]. A list of protein modifications observed
137 in the mass spectrometric analysis is shown in Supplemental Table 3.

138 Adenovirus penton proteins can spontaneously assemble into 12-pentamer subviral particles that
139 may serve as decoy pseudocapsids in vivo [14]. To investigate the possibility of a comparable
140 phenomenon in adomaviruses, codon-modified marbled eel adomavirus LO1-LO8 expression
141 plasmids were transfected individually into mammalian 293TT cells [15]. Optiprep
142 ultracentrifugation was used to separate virus-like particles from smaller solutes. A human
143 papillomavirus type 16 (HPV16) L1/L2 expression plasmid was used as a positive control for
144 virus-like particle (VLP) formation [16]. Cells transfected with adomavirus LO4, LO5, or LO7
145 expression constructs each showed robust production of particles that migrated into the core
146 fractions of Optiprep gradients (Figure 2), whereas cells transfected with LO1, LO2-3, LO6, or
147 LO8 alone did not show evidence of particle formation (Supplemental Figure 2). In co-
148 transfections of various combinations of LO genes, it was found that inclusion of LO6 inhibited
149 the formation of LO5 and LO7 particles but did not impair the formation of LO4 particles
150 (Supplemental Figure 3). This result suggests that LO6 is a minor virion component that directly
151 interacts with LO5 and LO7.

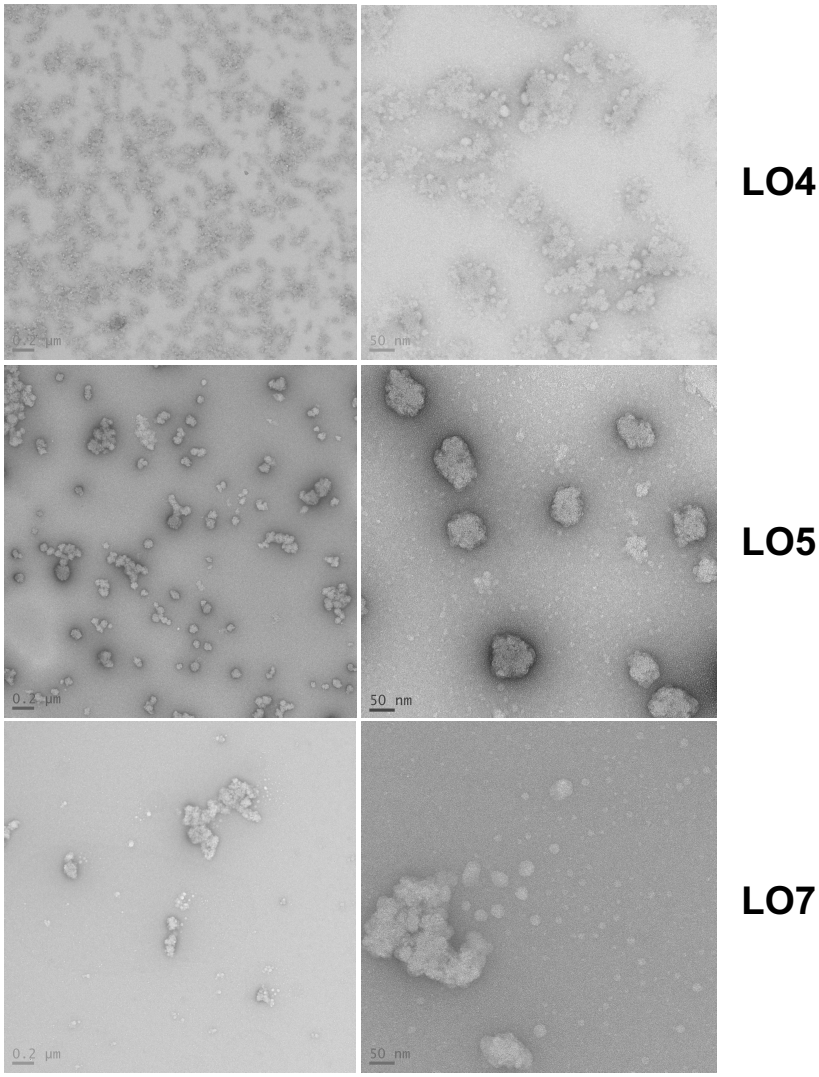


Figure 2. Negative stain EM of particles formed after expression of single LO proteins in mammalian cells. Mammalian 293TT cells were transfected with expression plasmids encoding the indicated marbled eel adomavirus proteins. Particles were extracted from cells by detergent lysis and purified by Optiprep ultracentrifugation. Scale bars are shown at the bottom left corner of each electron micrograph.

Individually expressed LO5 and LO7 particle preparations showed dsDNA signal in Quant-iT PicoGreen assays (Invitrogen), indicating the presence of nuclease-resistant encapsidated DNA within the purified particles. Optiprep-purified particle preparations from cells co-transfected with LO4, LO5, and LO7 were subjected to an additional round of nuclease digestion with salt-tolerant Benzonase endonuclease (Sigma) followed by agarose gel filtration to remove the nuclease and any residual digested DNA fragments. The gel filtered particles reproducibly contained roughly seven nanograms of DNA per microgram of total protein, confirming the presence of nuclease-resistant nonspecific cellular DNA within the particles.

161 Taken together, these results support the hypothesis that LO4, LO5, LO6, LO7, and LO8 are
162 syntenic homologs of adenovirus pIX, penton, pVI:pX, hexon, and adenain, respectively.
163

164 **Detection of additional adomaviruses**

165 In an effort to discover adomaviruses that infect animals amenable to laboratory study, we
166 performed sequencing of aquarium fish specimens and pursued data mining of public sequence
167 repositories. Post-mortem metagenomics analysis of an aquarium-bred Amazon red discus
168 cichlid (*Symphysodon discus*) exhibiting lethargy and inflamed skin lesions revealed a circular
169 ~20 kb adomavirus genome (Figure 3). Histopathological analysis of skin lesions from the discus
170 specimen showed no evidence of intranuclear inclusions or other obvious histopathology.

171 Another adomavirus from an apparently healthy green arowana (*Scleropages formosus*) was first
172 identified as a set of short NCBI Whole-Genome Shotgun (WGS) contigs with similarity to
173 Japanese eel adomavirus. A complete adomavirus genome was characterized by Sanger
174 sequencing of overlapping PCR products using DNA left over from the original fin snip used for
175 the WGS project [17].

176 A bioinformatics pipeline was developed to probe the SRA. Datasets that appeared to be rich in
177 adomavirus-like sequences based on translated nucleotides were subjected to de novo assembly.
178 This approach resulted in the identification of 10 complete or nearly complete adomavirus
179 genomes representing distinct species (Supplemental Table 1, Figure 3). Notably, two
180 adomaviruses were identified from genome sequencing datasets for farmed tilapia (*Oreochromis*
181 *niloticus*), which represent a $9 billion per year global aquaculture industry. Adomavirus-like
182 fragments were also detected in a brook salamander (*Calotriton asper*) and two closely related
183 species of anole lizard (genus *Anolis*). Our pipeline did not have adequate throughput to
184 comprehensively search the much larger volume of SRA data for mammals and birds.

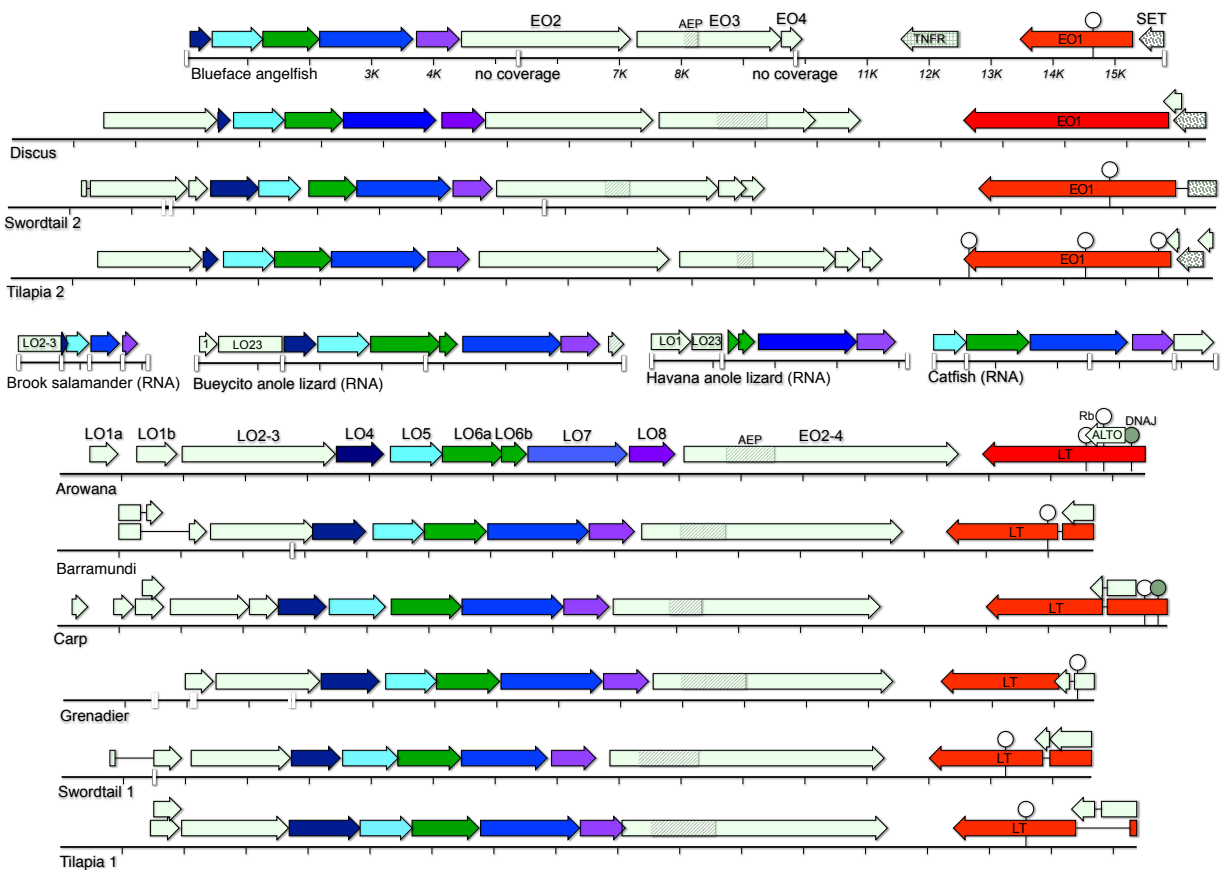


Figure 3: Picture maps of previously unknown adomaviruses. The figure uses the same color scheme as Figure 1. Alpha adomaviruses are defined by the presence of an adoma-specific superfamily 3 helicase (S3H) gene, designated EO1. Beta adomaviruses instead encode a polyomavirus large T antigen (LT). In many cases, repetitive or GC-rich patches (particularly near the 3' end of the LO2-3 ORF) could not be resolved using available short-read datasets. Coverage gaps are represented as white bars. A poorly conserved set of accessory genes upstream of beta adomavirus EO1 genes show varying degrees of similarity to the S-adenosyl methionine-binding pocket of cellular SET proteins, which function as histone lysine methyltransferases. Adomavirus SET homologs are highly divergent from all previously described eukaryotic and viral SET genes. The same is true for adomavirus EO2-4 and EO3 segments that encode homologs of the catalytic small subunit of archaeal eukaryotic primases (AEPs). Supplemental Table 1 lists virus accession numbers and Linnaean designations of host animals.

Adomavirus sequences can be divided into two groups based on their replicative S3H genes. A group we designate Alpha is defined by the presence of an EO1 replicase gene distantly similar to papillomavirus E1 proteins (Figure 4). Sequence and predicted structural similarities include a C-terminal S3H domain and a central HUH nicking endonuclease-like segment referred to as the origin binding domain [3,18,19] Kazlauskas, 2019 #117}. Members of adomavirus clade Beta encode proteins highly similar to the large T antigen (LT) proteins of fish-tropic polyomaviruses. The adomavirus LT-like proteins share a full range of familiar polyomavirus-like features, including an N-terminal DnaJ domain, a potential retinoblastoma-interaction motif (LXCXE or LXXLFD) [20], a nickase-like origin-binding domain, and a C-terminal S3H domain [21].

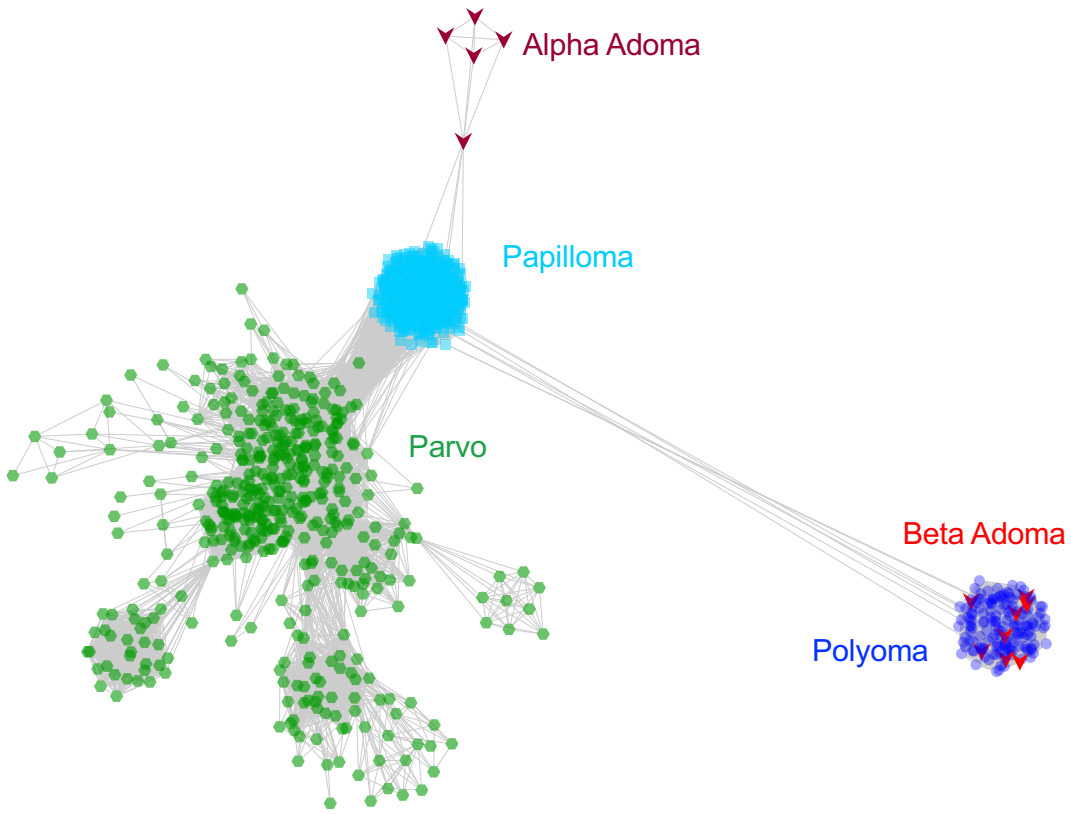


Figure 4: High stringency (E-10) sequence similarity network analysis of S3H proteins from indicated virus groups. An interactive Cytoscape file that incorporates virus names, accession numbers, and sequences is posted at https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoNetwork

It is thought that polyomaviruses and papillomaviruses tend to co-evolve with their animal hosts [22,23]. This is reflected in the fact that phylogeny within polyomavirus and papillomavirus genera recapitulates the phylogenetic relationships of host animals. Although adomavirus phylogenetic trees are currently sparse, the available data are consistent with adomavirus-host co-evolution similar to that of polyomaviruses and papillomaviruses (Supplemental Figure 4). For example, the phylogeny of adomavirus LO8 (adenain) proteins roughly correlates with the animal classes in which the sequences were found. In particular, Alpha adomavirus sequences observed in RNA datasets for brook salamander and anole lizards are more closely related to one another than to fish-associated adomaviruses. The results suggest that Alpha adomaviruses and Beta adomaviruses have been evolving as independent lineages for at least half a billion years (Supplemental Figure 4).

Discovery of adintoviruses

Database searches for adomaviruses revealed a large number of contigs uniting an adenain homolog, polintovirus-like penton- and hexon-like genes, and a retrovirus-like integrase (Figure 5, Supplemental Figure 5). We refer to this group of contigs as "adintoviruses." A hallmark adintovirus feature is a large protein with a characteristic N-terminal RNA-binding zinc finger-like domain, a central papain/OTU1-like domain, an endonuclease VII-like domain, and a C-terminal type B DNA polymerase (PolB) domain. A sequence similarity network analysis of

214 adintovirus PolB genes reveals two distinct clades (Figure 6). The predicted PolB proteins of a
215 lineage we designate Alpha show similarity to adenovirus PolB genes, while Beta adintovirus
216 PolB genes are distantly similar to bacteriophage PolB sequences. The appearance of two distinct
217 clades is recapitulated in network analyses of predicted adintovirus hexon and penton proteins,
218 suggesting the existence of two independent lineages.

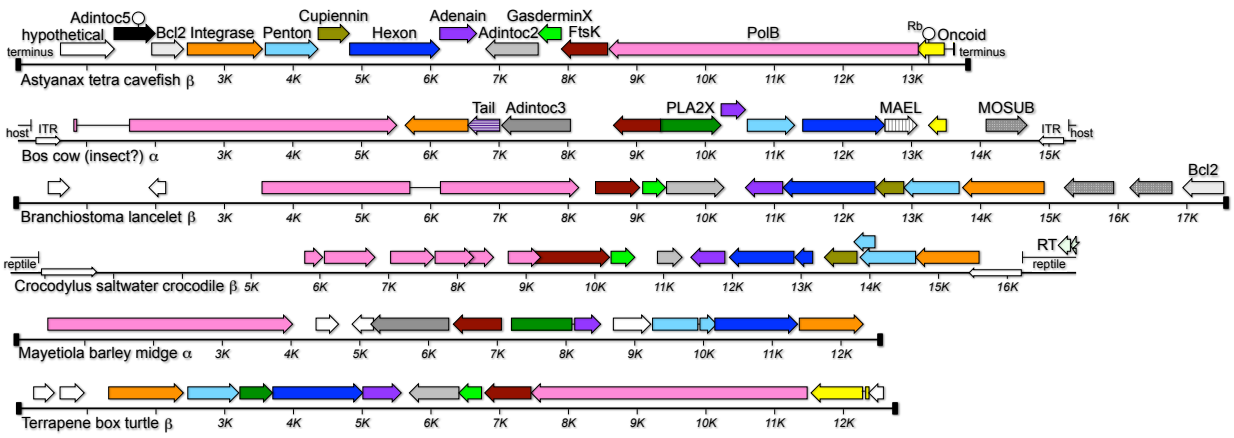


Figure 5: Picture maps of select adintovirus genomes. The figure uses the same color scheme as Figure 1. Supplemental Figure 5 shows additional examples of adintovirus genomes. Supplemental Table 1 lists accession numbers and host animal taxonomic designations.

219

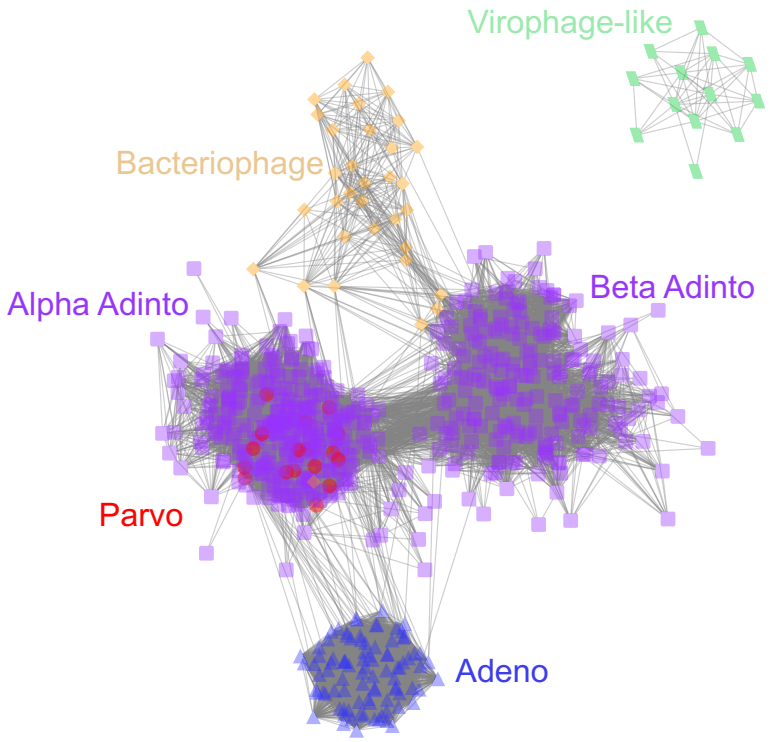


Figure 6: Moderate stringency (E-5) network analysis of PolB proteins. An interactive Cytoscape file that incorporates virus names, accession numbers, and sequences is posted at https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoNetwork

Apparently complete adintovirus genomes were found in high-throughput sequencing datasets for a wide variety of animal types (Supplemental Table 1 and https://ccrod.cancer.gov/confluence/display/LCOTF/AdintoAll). In many cases, the consensus of mapped reads ended abruptly, suggesting a linear, non-integrated viral genome. In some instances, such as a barley midge (*Mayetiola destructor*) dataset, a complete linear adintovirus genome could be assembled, but the dataset also contained a range of lower-coverage variant sequences, some of which were flanked by inverted terminal repeats (ITRs) contiguous with insect genomic DNA sequences. This observation suggests that the adintovirus integrase gene is functional and mediates integration events akin to those observed in virophages [9]. A subset of apparently integrated adintoviruses, including contigs assembled from various crocodile datasets, are heavily interspersed with frameshift and nonsense mutations (Figure 5). In many cases, adintovirus-like sequences appear to be adjacent to or infiltrated by endogenous retroelements. The results indicate that adintovirus sequences can exist as free viral minichromosomes, as integrated proviruses, or as endogenized degraded relics in animal germline DNA.

Many Alpha adintoviruses encode proteins similar to the phospholipase A2 (PLA2) domain of parvovirus VP1 virion proteins. The adintovirus PLA2 homologs typically include a C-terminal domain with predicted structural similarity to adenovirus pX. Adintovirus "PLA2X" proteins also show predicted structural similarity to melittin and secreted PLA2 proteins found in bee and snake venom. The fact that venom melittin and PLA2 act in concert [24] raises the speculative

240 hypothesis that some animal venom proteins might share common ancestry with membrane-
241 active viral capsid proteins.

242 Some adintoviruses encode homologs of the C-terminal regulatory domain of gasdermins, a
243 group of pore-forming proteins that serve as executioners in pyroptosis (a form of inflammatory
244 programmed cell death) [25]. Adintovirus homologs of a spider venom protein known as cupiennin
245 were also observed. Other predicted protein sequences were conserved among various
246 adintoviruses but did not show clear hits in BLAST or HHpred searches. We assigned these
247 groups of adinto-conserved unknown proteins numbered "Adintoc" names.

248 All classes of small DNA tumor viruses encode proteins harboring motifs, such as LXCXE, that
249 that are known to engage cellular retinoblastoma (Rb) and related tumor suppressor proteins [2].
250 Adenovirus E1A, papillomavirus E7, polyomavirus LT, and parvovirus NS3 oncoproteins
251 typically encode the Rb-binding motif just upstream of a consensus casein kinase 2 acceptor
252 motif ((ST)XX(DE)). Some oncogenes, such as E1A, encode an additional conserved motif,
253 referred to as CR1, that resembles Rb pAB groove-interacting motifs
254 ((DEN)(LIMV)XX(LM)(FY)) [20,26]. In general, these predicted Rb-interacting motifs are adjacent
255 to potential zinc- or iron-sulfur-binding motifs (typically, paired CXXC). Open reading frames
256 encoding combinations of these motifs are frequently observed in adintovirus contigs
257 (Supplemental Figure 5). We refer to these predicted proteins as "oncoids" due to their
258 similarities to the known oncogenes of small DNA tumor viruses. Predicted adintovirus
259 homologs of anti-apoptotic proteins, such as Bcl2 and IAP, were also observed. Surprisingly,
260 oncoid-like genes were also observed in virophage-like polintoviruses of unicellular organisms
261 (Supplemental Figure 6). These observations raise the hypothesis that the oncogenes of modern
262 small DNA tumor viruses descended from accessory genes that first arose in viruses of early
263 eukaryotes.

264 Adintoviruses were detected in diverse chordate datasets ranging from lancelets (*Branchiostoma*
265 *belcheri*) to box turtles (*Terrapene mexicana triunguis*). A seemingly complete adintovirus
266 genome was found in a transcriptomic dataset for a Mexican blind tetra cavefish (*Astyanax*
267 *mexicanus*). Adintovirus transcripts were most abundant in cavefish head, kidney, and intestine
268 samples and least abundant in muscle and whole embryo samples (Supplemental Table 4). In
269 contrast to the lancelet and box turtle datasets, which showed a high degree of intra-host
270 adintovirus sequence diversity at varying abundance, the cavefish adintovirus sequences were
271 relatively uniform. The observation is consistent with possible bottlenecking of a single
272 adintovirus strain into the sampled population of aquarium fish. Cavefish could provide a
273 tractable small vertebrate model for adintovirus infection.

274 The only obvious examples of adintoviruses in mammal datasets were found in a PacBio-based
275 WGS survey of bovine lung tissue. Sequences outside the ITRs were highly diverse but in a few
276 instances showed BLASTN similarity to genomic DNA sequences of various beetles, including
277 *Tribolium castaneum* (a flour beetle that commonly infests cattle feed). Similarly, two integrated
278 adintoviruses found in a genomic dataset for olive trees (*Olea europaea*) showed insect-like
279 sequences outside the ITRs (Supplemental Figure 5). Adintovirus sequences found in some
280 datasets thus appear to represent environmental contaminants, as opposed to infection of the
281 organism that was the subject of the sequencing effort.

282 **Identification of "missing link" parvoviruses**

283 Sequence queries in GenBank nr, WGS, and TSA databases revealed a large and highly diverse
284 range of parvovirus NS1-like genes in all major animal phyla. With the exception of parvovirus-
285 like contigs found in datasets for *Abeoforma whisleri* and *Ichthyophonus hoferi* (unicellular
286 eukaryotes that are closely related to animals), parvovirus-like sequences were not convincingly
287 detected in datasets for non-animal eukaryotes. Interestingly, the *Abeoforma* contig encodes a
288 protein with similarity (HHPred probability 82%) to the reverse transcriptase domain of group II
289 intron maturases (Figure 7).

290 BLAST searches using Alpha adintovirus PolB sequences consistently returned high-likelihood
291 matches to the PolB genes of an emerging group of bipartite parvoviruses referred to as
292 bidnaviruses [27]. SRA searches for additional examples of bidna-class PolB genes revealed
293 bidnaparvovirus-like contigs in datasets for false wolf spider (*Tengella perfuga*) and Tasmanian
294 devil (*Sarcophilus harrisii*) feces. The detected PolB genes exhibit predicted structural similarity
295 OTU1 (a hallmark feature of adintovirus PolB genes).

296 The bidnparvovirus-like genes observed in the false wolf spider and Tasmanian devil feces
297 datasets each occupied a single contig. In previously reported bidnaparvoviruses, the ITRs of
298 each of the two genome segments are identical. We therefore probed the spider and devil datasets
299 for examples of additional contigs with termini similar to the ITRs of the initially observed
300 contigs. The datasets were also searched for contigs encoding proteins similar to previously
301 reported bidnaparvovirus proteins. Second segments were not detected, suggesting that the
302 viruses are monopartite. We suggest the name "monodnaparvovirus" for this class of sequences.

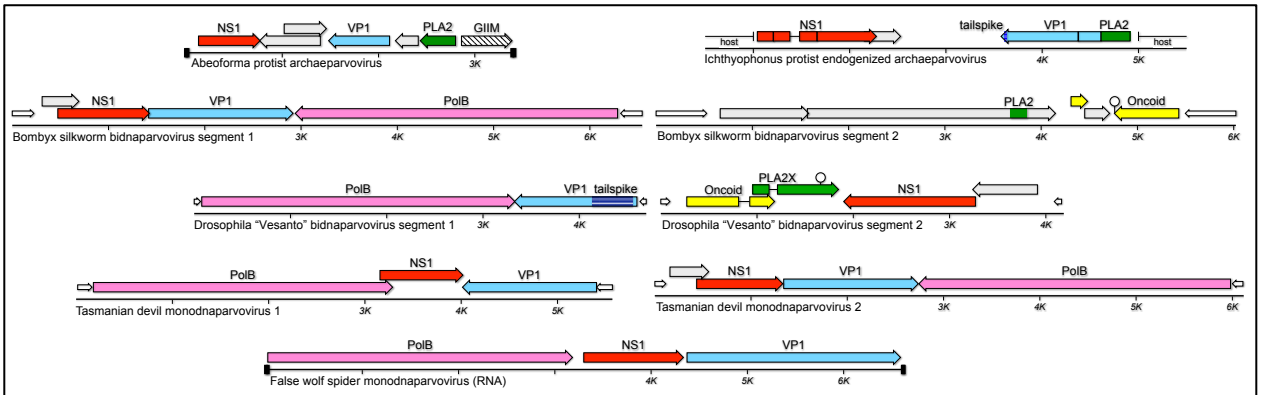


Figure 7. Parvovirus genome maps. Figure uses the same color scheme as Figure 1. GIIM indicates similarity to group II intron maturases. Accession numbers are listed in Supplemental Table 1.

303

304 **Analysis of S3H genes**

305 Polyomavirus LT, papillomavirus E1, and adomavirus EO1 replicase proteins have an origin-
306 binding domain that shows clear structural similarity to the HUH "nickase" endonuclease
307 domain of parvovirus NS1 proteins [3,8,18]. The replicase genes of CRESS viruses share the same
308 HUH-S3H domain architecture. Many CRESS viruses encode an ORF "overprinted" in the +1
309 frame of the Rep ORF and similar overprinted ORFs have been observed in parvovirus NS1,
310 polyomavirus LT, and papillomavirus E1 genes [28,29].

311 We did not detect examples of adintovirus-specific S3H genes. Although many virophage-class
312 polintoviruses encode proteins with S3H domains [30], our searches confirm that the S3H proteins
313 of virophage-like viruses lack detectable HUH endonuclease-like domains and lack overprinted
314 ORFs. Furthermore, virophage-class S3H proteins did not cluster with small DNA tumor virus
315 S3H proteins in sequence similarity network analyses. These observations support the standard
316 view that CRESS, polyoma, papilloma, and parvovirus replicase genes share more recent
317 common ancestry with one another than with the S3H genes of other known virus lineages [3,8].

318

319 **Discussion**

320 In this report, we began by identifying the hexon and penton genes of the emerging adomavirus
321 family. We demonstrated that the adomavirus LO4-8 array is homologous to the adenovirus late
322 gene array that encodes most of the virion structural proteins. The ability of adomavirus virion
323 proteins to assemble into irregular "virus-like" particles should facilitate the development of
324 recombinant vaccines against adomaviruses.

325 We identified two distinct lineages of adomaviruses that appear to have been independently co-
326 evolving with their vertebrate hosts for at least half a billion years (Supplemental Figure 4).
327 Within both of these adomavirus lineages, there are members associated with commercially
328 important fish species. Surprisingly, three closely related Alpha adomavirus-like segments were
329 observed in RNA sequencing datasets for terrestrial vertebrates. It will be important to apply
330 emerging higher-throughput search algorithms, such as Mash Screen [31], to exhaustively search
331 for adomaviruses in deep sequencing datasets for other terrestrial vertebrates, including humans.

332 Our searches for adomaviruses serendipitously revealed an apparently older lineage of animal-
333 tropic viruses that we name adintoviruses. Adintoviruses encode a number of genes that appear
334 to be homologs of membrane-active proteins found in animal venom. These include bee and
335 snake venom PLA2 and melittin, as well as a spider venom protein called cupiennin. In some
336 cases, these proteins also score as weakly similar to bacterial colicin proteins or bacteriophage
337 holin proteins, which play a role in disrupting bacterial membranes during the infectious entry
338 process. We speculate that some types of animal venom genes may be endogenized membrane-
339 active virion protein genes. It also appears that adintoviruses share relatively recent common
340 ancestry with some segments of bracoviruses (Supplemental Figure 5), a group of endogenized
341 virus-like elements that parasitoid wasps inject alongside their eggs [32].

342 Adomaviruses and adintoviruses have a number of features that could make them useful as
343 recombinant gene transfer vectors. Their genome size is substantially larger than commonly used
344 parvoviral, lentiviral, and papillomaviral vectors but, in contrast to adenoviruses, their genomes
345 are small enough to be manipulated in the setting of standard plasmids. An intriguing feature of
346 the adintovirus integrase gene is the presence of a predicted chromodomain that, in LTR
347 retrotransposons, is believed to influence integration site specificity [33]. Adintoviruses could
348 theoretically provide a useful alternative to retroviral vectors.

349 It has generally been assumed that the functionally similar oncogenes found in adenoviruses,
350 papillomaviruses, parvoviruses, and polyomaviruses arose through convergent evolution or
351 through horizontal gene transfer between virus families [2]. Although small DNA tumor virus

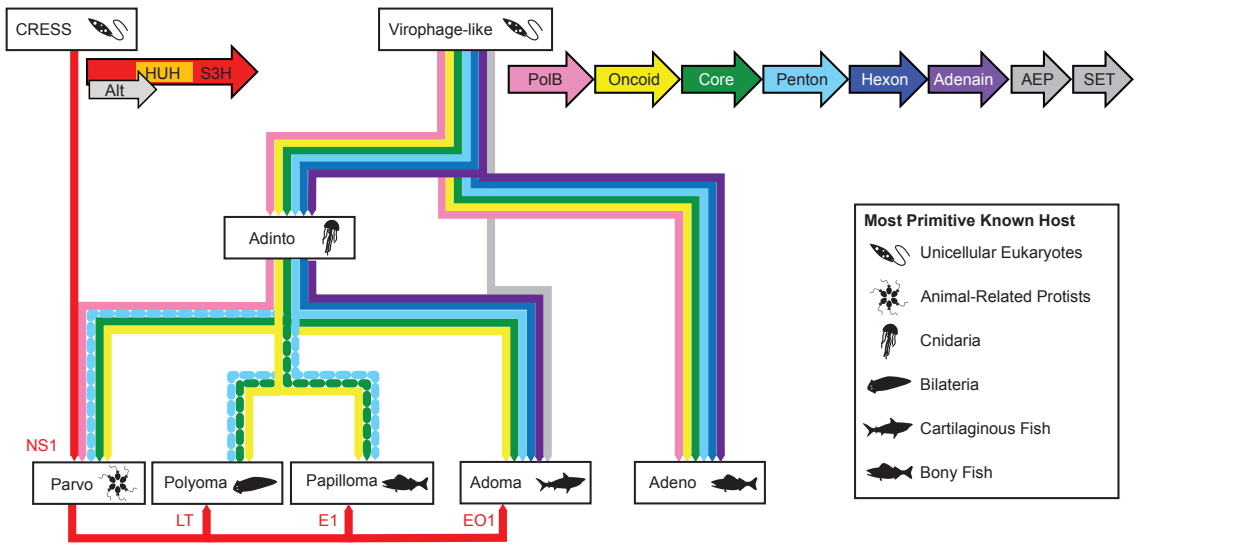
352 oncogenes show extremely low overall sequence similarity, they can be roughly defined based
353 on the presence of short linear motifs. Many adintoviruses encode candidate "oncoid" proteins
354 with these motifs. Unexpectedly, it appears that some virophage-class polintoviruses of
355 unicellular organisms also encode oncogene-like proteins (Supplemental Figure 6). These
356 observations raise the hypothesis that the cell growth-regulating oncogenes of modern small
357 DNA tumor viruses might have descended from a class of viral accessory genes that evolved
358 before the appearance of multicellular animals.

359 Although the current study indicates clear homologies between adenovirus, adomavirus, and
360 adintovirus virion proteins, the origins of polyomavirus and papillomavirus virion proteins
361 remain unclear. It has previously been suggested that the major capsid proteins the smaller tumor
362 virus families might have arisen from the capsid proteins of CRESS and/or ssRNA viruses [3,8].
363 This would have involved a complex shift from the simple T=1 icosahedral geometry of CRESS
364 virions, in which the strands of the core ß-jellyroll of the major capsid protein are parallel to the
365 lumen of the virion, to the more complex T=7 perpendicular jellyroll organization seen in
366 papillomaviruses and polyomaviruses [1]. The process would also have involved a transition from
367 encapsidation of a naked ssDNA viral genome to encapsidation of a chromatinized dsDNA
368 genome. Since CRESS viruses are not known to encode minor capsid proteins, descent of
369 papillomaviruses and papillomaviruses would also have involved the acquisition of the hallmark
370 L2 and VP2 minor capsid proteins from unknown sources. CRESS viruses typically have
371 genomes less than 3 kb in size, while papillomaviruses and polyomaviruses are typically 5-8 kb
372 in size. Although a few CRESS viruses with 5-9 kb genomes have recently been reported, the
373 larger CRESS virus genomes each encode multiple parallel-jellyroll capsid protein paralogs or
374 unrelated non-jellyroll filamentous virion proteins [34]. There thus does not appear to be a clear
375 precedent for papillomavirus/polyomavirus-like virion architecture among known CRESS
376 viruses. We propose a simpler model in which the pentameric papillomavirus and polyomavirus
377 major capsid proteins were derived from the perpendicular single-jellyroll penton proteins of
378 adintovirus-like ancestors and the minor capsid proteins were derived from a penton-associated
379 minor virion protein. The observation that the penton protein of marbled eel adomavirus
380 associates can assemble into roughly spherical ~50 nm structures carrying non-specifically
381 encapsidated dsDNA raises the possibility that minimal evolutionary adaptations would be
382 required to construct a structure capable of housing a polyomavirus or papillomavirus genome.
383 This hypothesis could also be extended to the splayed-jellyroll capsomers of parvoviruses. One
384 line of support for this idea is the observation that parvovirus virion proteins typically include an
385 N-terminal domain closely resembling adintovirus PLA2X proteins. Similarly, a small subset of
386 parvoviruses appear to encode capsid protein domains with distant similarity to predicted
387 adenovirus tailspike-like proteins [35,36], which some adintoviruses appear to encode.

388 To further explore the hypothesis that papillomavirus, polyomavirus, and parvovirus virion
389 proteins were derived from adintovirus-like ancestors, extremely low stringency network
390 analyses were performed for various classes of major and minor capsid proteins. The analysis
391 rests on the assumption that a small subset of members of each virus family might have better
392 preserved the primary sequence of a common ancestor, which would be revealed by BLAST-
393 detectable linkages between a few members of different clusters. Aside from the previously
394 described distant similarity between the parvovirus PLA2 domain and the Cap genes of small
395 subset of non-animal CRESS viruses [37,38], no clear connections were observed between groups of
396 small DNA tumor viruses and CRESS virus capsid proteins. The low-stringency analysis does,

397 however, suggest possible similarities between papillomavirus, polyomavirus, and parvovirus
398 major capsid proteins and the penton proteins of adenoviruses, adomaviruses, and adintoviruses
399 (Supplemental Figure 7). Possible distant similarity was also observed for some combinations of
400 minor capsid proteins (Supplemental Figure 8).

401 A framework that could account for the apparent similarities between extant small DNA tumor
402 viruses is presented in Figure 8. A testable prediction of this model is that adomavirus LO6 and
403 adintovirus PLA2X, cupiennin, and gasdermin proteins will, like the papillomavirus and
404 polyomavirus minor capsid proteins, be found to directly interact with the luminal surface of
405 penton proteins in the assembled virion. We also predict that the fold of polyomavirus VP2 and
406 papillomavirus L2 proteins may be found to resemble inferred minor capsid proteins of
407 adintoviruses. The model further predicts that adenovirus pX, in addition to its proposed role in
408 condensing adenovirus genomes [11], participates in destabilization of host cell membranes during
409 the infectious entry process in a manner similar to that of papillomavirus L2 proteins [39,40].
410 Another prediction is that the penton proteins of adintoviruses can, like marbled eel adomavirus
411 penton protein, form spherical 50 nm structures that contain nuclease-resistant DNA.



412 **Figure 8: A model for gene flow during small DNA tumor virus evolution.** Dotted lines are hypothetical.

413 At the outset of this study, adintoviruses were not recognizable as viruses. For example,
414 comprehensive searches for adventitious viruses in the commonly used *Spodoptera frugiperda*
415 cell line SF9 have misclassified adintoviruses as transposons [41,42]. An important implication is
416 that there may be additional unrecognized families of animal viruses hiding in plain sight in
417 sequence databases. To address this problem, it will be important to develop higher throughput
418 methods for more sensitive structure-guided searches and for understanding which combinations
419 predicted homologs co-occupy single contigs. It will be especially important to apply high
420 throughput annotation pipelines, such as Cenote-Taker [34], that generate human-readable
421 GenBank compliant maps. Deposition of these annotated new viral genomes in publicly
422 accessible databases will be critical for expanding our understanding of the animal virome.

423

424 Materials and Methods

425 **Sample Collection and cell culture**

426 A red discus cichlid (*Symphysodon discus*) was purchased at a pet shop in Gainesville, Florida.
427 The fish was moribund and showed erythematous skin lesions. Propagation of the discus
428 adomavirus in cell culture was attempted by overlaying skin tissue homogenates on Grunt Fin
429 (GF) and Epithelioma Papulosum Cyprini cell lines (ATCC). Neither cytopathic effects nor
430 qPCR-based detection of viral replication were observed during two blind passages of 14 days
431 each.

432 Dr. Chiu-Ming Wen generously provided EK-1 cells (a Japanese eel kidney line) infected with
433 the Taiwanese marbled eel adomavirus [6]. The virus was propagated by inoculation of
434 supernatants from the infected culture into uninfected EK-1 cells cultured at room temperature in
435 DMEM with 10 % fetal calf serum. Human embryonic kidney-derived 293TT cells were cultured
436 as previously described [15].

437 **Viral genome sequencing**

438 For the discus adomavirus, total DNA was extracted from a skin lesion and subjected to deep
439 sequencing. Marbled eel adomavirus virions were purified from lysates of infected EK-1 cells
440 using Optiprep gradient ultracentrifugation [43]. DNA extracted from Optiprep gradient fractions
441 was subjected to rolling circle amplification (RCA, TempliPhi, GE Health Sciences). The
442 marbled eel adomavirus RCA products and discus total DNA were prepared with a Nextera XT
443 DNA Sample Prep kit and sequenced using the MiSeq (Illumina) sequencing system with 2 ×
444 250 bp paired-end sequencing reagents. In addition, the marbled eel adomavirus RCA product
445 was digested with *AclI* and *EcoRI* restriction enzymes and the resulting early and late halves of
446 the viral genome were cloned separately into the *AclI* and *EcoRI* restriction sites of pAsylum+.
447 The sequence of the cloned genome was verified by a combination of MiSeq and Sanger
448 sequencing. The clones will be made available upon request.

449 For the arowana adomavirus, overlapping PCR primers were designed based on WGS accession
450 numbers LGSE01029406, LGSE01031009, LGSE01028643, LGSE01028176, and
451 LGSE01030049 [17]. PCR products were subjected to primer walking Sanger sequencing.

452 **Marbled eel adomavirus transcript analysis, late ORF expression, and virion purification**

453 RNAseq reads reported by Wen et al [6] were aligned to the marbled eel adomavirus genome using
454 HISAT2 version 2.0.5 [44] with the following options: "--rna-strandness FR --dta --no-mixed --no-
455 discordant". Integrated Genome Viewer (IGV) version 2.4.9 [45] was used to determine splice
456 junctions and their depth of coverage. Additional validation was performed by visual inspection
457 using CLC Genomics Workbench 12.

458 Codon-modified expression constructs encoding the marbled eel adomavirus LO1-LO8 proteins
459 were designed according to a modified version of a previously reported algorithm
460 (https://github.com/BUCK-LCO-NCI/Codmod_as_different_as_possible)[46]. 293TT cells were
461 transfected with LO expression constructs for roughly 48 hours. Cells were lysed in a small
462 volume of PBS with 0.5% Triton X-100 or Brij-58 and Benzonase Dnase/Rnase (Sigma)[47]. After

463 one hour of maturation at neutral pH, the lysate was clarified at 5000 x g for 10 min. The
464 clarified lysate was loaded onto a 15-27-33-39-46% Optiprep gradient in PBS with 0.8 M NaCl.
465 Gradient fractions were collected by bottom puncture of the tube and screened by PicoGreen
466 dsDNA stain (Invitrogen), BCA, or SDS-PAGE analysis. Electron microscopic analysis was
467 performed by spotting several microliters of Optiprep fraction material (or, in some instances,
468 particles exchanged out of Optiprep using agarose gel filtration) onto carbon support film on
469 Cu/Ni/Au grids, followed by staining with 0.5% uranyl acetate.
470

471 **Discovery of viral sequences in NCBI databases**

472 Compilations of sequences use in this study are posted at the following link:
473 https://ccrod.cancer.gov/confluence/display/LCOTF/Adinto. Papillomavirus E1 sequences were
474 downloaded from PaVE https://pave.niaid.nih.gov [48]. Polyomavirus LT sequences were
475 downloaded from PyVE https://ccrod.cancer.gov/confluence/display/LCOTF/Polyomavirus [23].
476 Parvovirus NS1 proteins and S3H proteins of virophage-like viruses were compiled from
477 multiple databases, including RefSeq, WGS, and TSA, using TBLASTN searches. Adenovirus,
478 virophage, and bacteriophage PolB sequences were downloaded from GenBank nr using
479 DELTA-BLAST searches [49] with alpha or beta adintovirus PolB proteins as bait.

480 SRA datasets for fish, amphibians, and reptiles were searched using DIAMOND [50] or NCBI
481 SRA Toolkit (http://www.ncbi.nlm.nih.gov/books/NBK158900/) in TBLASTN mode using
482 adomavirus protein sequences as the subject database or query, respectively. Reads with
483 similarity to the baits were collected and subjected to BLASTX searches against a custom library
484 of viral proteins representing adomaviruses and other small DNA tumor viruses. SRA datasets of
485 interest were subjected to de novo assembly using the SPAdes suite [51,52] or Megahit [53,54]. Contigs
486 encoding virus-like proteins were identified by TBLASTN searches against adomavirus protein
487 sequences using Bowtie [55]. The candidate contigs were validated using the CLC Genomics
488 Workbench 12 align to reference function.

489 Adintoviruses were extracted from WGS and TSA datasets using TBLASTN searches with S3H
490 genes of various small DNA tumor viruses or the PolB, adenain, or PLA2 genes of Nephila orb-
491 weaver spider alpha adintovirus GFKT014647032. Predicted protein sequences were
492 automatically extracted from contigs of interest using getorf (http://bioinfo.nhri.org.tw/cgi-
493 bin/emboss/getorf) [56]. Proteins were clustered using EFI-EST (https://efi.igb.illinois.edu/efi-
494 est/)[57,58] and displayed using Cytoscape v3.7.1 [59]. Multiple sequence alignments were
495 constructed using MAFFT (https://toolkit.tuebingen.mpg.de/#/tools/mafft). Individual or aligned
496 protein sequences were subjected to HHpred searches
497 (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) [60] of PDB, Pfam-A, NCBI Conserved Domains,
498 and PRK databases.

499 Contigs were annotated using Cenote-Taker [34] with an iteratively refined library of conserved
500 adintovirus protein sequences. Maps were drawn using MacVector 17 software.

501 The phylogenetic analysis shown in Supplemental Figure 4 was performed using Phylogeny.fr [61].
502 Host animal divergence times were obtained from TimeTree of Life http://www.timetree.org [62].

503 **Mass Spectrometry**

504 Optiprep-purified marbled eel adomavirus virions were precipitated with trichloroacetic acid. A
505 1 ml sample was treated with 100 µl of 0.15% deoxycholic acid and incubated at room
506 temperature for 10 minutes. 100 µl of 100% TCA was then added and the sample was vortexed
507 and incubated on ice for 30 minutes. Following the incubation, the sample was centrifuged at
508 10,000 x g for 10 minutes at 4°C. The supernatant was removed, and the remaining pellet was
509 washed with ice-cold acetone to remove residual TCA. The protein pellet was solublized with
510 NuPAGE Sample Buffer + 5% BME (Sigma) and run on a 10-12% Bis-Tris MOPS gel
511 (Thermo). The protein bands were visualized using InstantBlue (Expedeon). Thirteen gel bands
512 were individually excised and placed into 1.5 ml Eppendorf tubes. The gel bands were sent to the
513 National Cancer Institute in Fredrick, Maryland where they were de-stained, digested with
514 trypsin, and processed on a Thermo Fisher Q Exactive HF Mass Spectrometer. Thermo Proteome
515 Discoverer 2.2 software was used for initial protein identification. The uninterpreted mass
516 spectral data were also searched against *Anguilla* proteins (Swiss-Prot and TrEMBL database
517 containing 105,268 proteins), Bos taurus proteins (Swiss-Prot and TrEMBL database containing
518 48,288 proteins), a common contaminants database (cRAPome), and translated marbled eel
519 adomavirus ORFs. Further analysis was conducted using Protein Metrics Biopharma software to
520 identify modifications missed in initial analyses.

521 **Ethics Statement**

522 All animal tissue samples were received as diagnostic specimens collected for pathogen testing
523 and disease investigation purposes.

524 **Data Availability**

525 GenBank accession numbers for sequences deposited in association with this study are
526 MF946549, MF946550, MH282863, and BK010888-BK010894. NCBI BankIt submissions
527 2250743-2256014 (see Supplemental Table 1) are presently awaiting assignment of accession
528 numbers. Compilations of all sequences used in this study are posted at
529 https://ccrod.cancer.gov/confluence/display/LCOTF/Adinto

530 **Acknowledgments**

531 The authors are indebted to Eugene Koonin and Natalya Yutin for their generous guidance and
532 for the spirited discussions that inspired us to pursue this study. We are particularly grateful to
533 them for sharing their observation that adomaviruses encode a recognizable adenain homolog
534 and their discovery of adomavirus sequences in the arowana WGS datasets. The authors are also
535 grateful to Lisa Jenkins for her extensive technical guidance on analyzing mass spectrometric
536 data. We thank Karl Münger for useful discussions, including advice about oncogene sequence
537 motifs.

538 **References**

539

540

541 1 Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular
542 ancestors. *Proc Natl Acad Sci U S A* **114**, E2401-E2410, doi:10.1073/pnas.1621061114
543 (2017).
544 2 de Souza, R. F., Iyer, L. M. & Aravind, L. Diversity and evolution of chromatin proteins
545 encoded by DNA viruses. *Biochim Biophys Acta* **1799**, 302-318,
546 doi:10.1016/j.bbagrm.2009.10.006 (2010).
547 3 Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of
548 eukaryotes: The ultimate modularity. *Virology* **479-480**, 2-25,
549 doi:10.1016/j.virol.2015.02.039 (2015).
550 4 Mizutani, T. *et al.* Novel DNA virus isolated from samples showing endothelial cell
551 necrosis in the Japanese eel, Anguilla japonica. *Virology* **412**, 179-187,
552 doi:10.1016/j.virol.2010.12.057 (2011).
553 5 Okazaki, S. *et al.* Detection of Japanese eel endothelial cells-infecting virus in Anguilla
554 japonica elvers. *J Vet Med Sci* **78**, 705-707, doi:10.1292/jvms.15-0515 (2016).
555 6 Wen, C. M., Chen, M. M., Wang, C. S., Liu, P. C. & Nan, F. H. Isolation of a novel
556 polyomavirus, related to Japanese eel endothelial cell-infecting virus, from marbled eels,
557 Anguilla marmorata (Quoy & Gaimard). *J Fish Dis*, doi:10.1111/jfd.12423 (2015).
558 7 Dill, J. A., Camus, A. C., Leary, J. H. & Ng, T. F. F. Microscopic and Molecular
559 Evidence of the First Elasmobranch Adomavirus, the Cause of Skin Disease in a Giant
560 Guitarfish, Rhynchobatus djiddensis. *MBio* **9**, doi:10.1128/mBio.00185-18 (2018).
561 8 Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of
562 prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal
563 plasmids. *Nat Commun* **10**, 3425, doi:10.1038/s41467-019-11433-0 (2019).
564 9 Fischer, M. G. & Hackl, T. Host genome integration and giant virus-induced reactivation
565 of the virophage mavirus. *Nature* **540**, 288-291, doi:10.1038/nature20593 (2016).
566 10 Logan, J. & Shenk, T. Adenovirus tripartite leader sequence enhances translation of
567 mRNAs late after infection. *Proc Natl Acad Sci U S A* **81**, 3655-3659 (1984).
568 11 Nemerow, G. R., Stewart, P. L. & Reddy, V. S. Structure of human adenovirus. *Curr*
569 *Opin Virol* **2**, 115-121, doi:10.1016/j.coviro.2011.12.008 (2012).
570 12 Moyer, C. L., Besser, E. S. & Nemerow, G. R. A Single Maturation Cleavage Site in
571 Adenovirus Impacts Cell Entry and Capsid Assembly. *J Virol* **90**, 521-532,
572 doi:10.1128/JVI.02014-15 (2015).
573 13 Ruzindana-Umunyana, A., Imbeault, L. & Weber, J. M. Substrate specificity of
574 adenovirus protease. *Virus Res* **89**, 41-52 (2002).
575 14 Vragniau, C. *et al.* Studies on the Interaction of Tumor-Derived HD5 Alpha Defensins
576 with Adenoviruses and Implications for Oncolytic Adenovirus Therapy. *J Virol* **91**,
577 doi:10.1128/JVI.02030-16 (2017).
578 15 Buck, C. B., Pastrana, D. V., Lowy, D. R. & Schiller, J. T. Efficient intracellular
579 assembly of papillomaviral vectors. *J Virol* **78**, 751-757 (2004).
580 16 Buck, C. B., Thompson, C. D., Pang, Y. Y., Lowy, D. R. & Schiller, J. T. Maturation of
581 papillomavirus capsids. *J Virol* **79**, 2839-2846, doi:10.1128/JVI.79.5.2839-2846.2005
582 (2005).

583 17 Bian, C. *et al.* The Asian arowana (Scleropages formosus) genome provides new insights
584 into the evolution of an early lineage of teleosts. *Sci Rep* **6**, 24501,
585 doi:10.1038/srep24501 (2016).
586 18 Hickman, A. B., Ronning, D. R., Kotin, R. M. & Dyda, F. Structural unity among viral
587 origin binding proteins: crystal structure of the nuclease domain of adeno-associated
588 virus Rep. *Mol Cell* **10**, 327-337 (2002).
589 19 Iyer, L. M., Koonin, E. V., Leipe, D. D. & Aravind, L. Origin and evolution of the
590 archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural
591 insights and new members. *Nucleic Acids Res* **33**, 3875-3896, doi:10.1093/nar/gki702
592 (2005).
593 20 Gouw, M. *et al.* The eukaryotic linear motif resource - 2018 update. *Nucleic Acids Res*
594 **46**, D428-D434, doi:10.1093/nar/gkx1077 (2018).
595 21 An, P., Saenz Robles, M. T. & Pipas, J. M. Large T antigens of polyomaviruses: amazing
596 molecular machines. *Annu Rev Microbiol* **66**, 213-236, doi:10.1146/annurev-micro-
597 092611-150154 (2012).
598 22 Rector, A. *et al.* Ancient papillomavirus-host co-speciation in Felidae. *Genome biology* **8**,
599 R57, doi:10.1186/gb-2007-8-4-r57 (2007).
600 23 Buck, C. B. *et al.* The Ancient Evolutionary History of Polyomaviruses. *PLoS pathogens*
601 **12**, e1005574, doi:10.1371/journal.ppat.1005574 (2016).
602 24 Vogt, W., Patzer, P., Lege, L., Oldigs, H. D. & Wille, G. Synergism between
603 phospholipase A and various peptides and SH-reagents in causing haemolysis. *Naunyn*
604 *Schmiedebergs Arch Pharmakol* **265**, 442-454 (1970).
605 25 Dubois, H. *et al.* Nlrp3 inflammasome activation and Gasdermin D-driven pyroptosis are
606 immunopathogenic upon gastrointestinal norovirus infection. *PLoS pathogens* **15**,
607 e1007709, doi:10.1371/journal.ppat.1007709 (2019).
608 26 Pipas, J. M. Common and unique features of T antigens encoded by the polyomavirus
609 group. *J Virol* **66**, 3979-3985 (1992).
610 27 Krupovic, M. & Koonin, E. V. Evolution of eukaryotic single-stranded DNA viruses of
611 the Bidnaviridae family from genes of four other groups of widely different viruses. *Sci*
612 *Rep* **4**, 5347, doi:10.1038/srep05347 (2014).
613 28 Carter, J. J. *et al.* Identification of an overprinting gene in Merkel cell polyomavirus
614 provides evolutionary insight into the birth of viral genes. *Proc Natl Acad Sci U S A* **110**,
615 12744-12749, doi:10.1073/pnas.1303526110 (2013).
616 29 Canuti, M. *et al.* New Insight Into Avian Papillomavirus Ecology and Evolution From
617 Characterization of Novel Wild Bird Papillomaviruses. *Front Microbiol* **10**, 701,
618 doi:10.3389/fmicb.2019.00701 (2019).
619 30 Yutin, N., Shevchenko, S., Kapitonov, V., Krupovic, M. & Koonin, E. V. A novel group
620 of diverse Polinton-like viruses discovered by metagenome analysis. *BMC Biol* **13**, 95,
621 doi:10.1186/s12915-015-0207-4 (2015).
622 31 Ondov, B. D. *et al.* Mash Screen: High-throughput sequence containment estimation for
623 genome discovery. *bioRxiv*, 557314, doi:10.1101/557314 (2019).
624 32 Strand, M. R. & Burke, G. R. Polydnaviruses: Nature's Genetic Engineers. *Annu Rev*
625 *Virol* **1**, 333-354, doi:10.1146/annurev-virology-031413-085451 (2014).
626 33 Kordis, D. A genomic perspective on the chromodomain-containing retrotransposons:
627 Chromoviruses. *Gene* **347**, 161-173, doi:10.1016/j.gene.2004.12.017 (2005).

628 34 Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses.
629 555375, doi:10.1101/555375 %J bioRxiv (2019).
630 35 Penzes, J. J. *et al.* Molecular characterization of a lizard adenovirus reveals the first
631 atadenovirus with two fiber genes and the first adenovirus with either one short or three
632 long fibers per penton. *J Virol* **88**, 11304-11314, doi:10.1128/JVI.00306-14 (2014).
633 36 Menendez-Conejero, R. *et al.* Structure of a Reptilian Adenovirus Reveals a Phage
634 Tailspike Fold Stabilizing a Vertebrate Virus Capsid. *Structure* **25**, 1562-1573 e1565,
635 doi:10.1016/j.str.2017.08.007 (2017).
636 37 Xu, B. *et al.* Hybrid DNA virus in Chinese patients with seronegative hepatitis
637 discovered by deep sequencing. *Proc Natl Acad Sci U S A* **110**, 10264-10269,
638 doi:10.1073/pnas.1303744110 (2013).
639 38 Naccache, S. N. *et al.* The perils of pathogen discovery: origin of a novel parvovirus-like
640 hybrid genome traced to nucleic acid extraction spin columns. *J Virol* **87**, 11966-11977,
641 doi:10.1128/jvi.02323-13 (2013).
642 39 Aydin, I. *et al.* A central region in the minor capsid protein of papillomaviruses facilitates
643 viral genome tethering and membrane penetration for mitotic nuclear entry. *PLoS*
644 *pathogens* **13**, e1006308, doi:10.1371/journal.ppat.1006308 (2017).
645 40 Campos, S. K. Subcellular Trafficking of the Papillomavirus Genome during Initial
646 Infection: The Remarkable Abilities of Minor Capsid Protein L2. *Viruses* **9**,
647 doi:10.3390/v9120370 (2017).
648 41 Geisler, C. A new approach for detecting adventitious viruses shows Sf-rhabdovirus-
649 negative Sf-RVN cells are suitable for safe biologicals production. *BMC Biotechnol* **18**,
650 8, doi:10.1186/s12896-017-0412-z (2018).
651 42 Nandakumar, S., Ma, H. & Khan, A. S. Whole-Genome Sequence of the Spodoptera
652 frugiperda Sf9 Insect Cell Line. *Genome Announc* **5**, doi:10.1128/genomeA.00829-17
653 (2017).
654 43 Peretti, A., FitzGerald, P. C., Bliskovsky, V., Buck, C. B. & Pastrana, D. V. Hamburger
655 polyomaviruses. *J Gen Virol* **96**, 833-839, doi:10.1099/vir.0.000033 (2015).
656 44 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory
657 requirements. *Nature methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).
658 45 Robinson, J. T., Thorvaldsdottir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant
659 Review with the Integrative Genomics Viewer. *Cancer Res* **77**, e31-e34,
660 doi:10.1158/0008-5472.CAN-17-0337 (2017).
661 46 Pastrana, D. V. *et al.* Reactivity of human sera in a sensitive, high-throughput
662 pseudovirus-based papillomavirus neutralization assay for HPV16 and HPV18. *Virology*
663 **321**, 205-216 (2004).
664 47 Buck, C. B. & Thompson, C. D. Production of papillomavirus-based gene transfer
665 vectors. *Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al*
666 **Chapter 26**, Unit 26.21 (2007).
667 48 Van Doorslaer, K. *et al.* The Papillomavirus Episteme: a major update to the
668 papillomavirus sequence database. *Nucleic Acids Res* **45**, D499-d506,
669 doi:10.1093/nar/gkw879 (2017).
670 49 Boratyn, G. M. *et al.* Domain enhanced lookup time accelerated BLAST. *Biology direct*
671 **7**, 12, doi:10.1186/1745-6150-7-12 (2012).
672 50 Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
673 DIAMOND. *Nature methods* **12**, 59-60, doi:10.1038/nmeth.3176 (2015).

674 51 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to
675 single-cell sequencing. *Journal of computational biology : a journal of computational*
676 *molecular cell biology* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).
677 52 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
678 metagenomic assembler. *Genome research* **27**, 824-834, doi:10.1101/gr.213959.116
679 (2017).
680 53 Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-
681 node solution for large and complex metagenomics assembly via succinct de Bruijn
682 graph. *Bioinformatics* **31**, 1674-1676, doi:10.1093/bioinformatics/btv033 (2015).
683 54 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by
684 advanced methodologies and community practices. *Methods* **102**, 3-11,
685 doi:10.1016/j.ymeth.2016.02.020 (2016).
686 55 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature*
687 *methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
688 56 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open
689 Software Suite. *Trends Genet* **16**, 276-277 (2000).
690 57 Gerlt, J. A. *et al.* Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web
691 tool for generating protein sequence similarity networks. *Biochim Biophys Acta* **1854**,
692 1019-1037, doi:10.1016/j.bbapap.2015.04.015 (2015).
693 58 Zallot, R., Oberg, N. O. & Gerlt, J. A. 'Democratized' genomic enzymology web tools for
694 functional assignment. *Curr Opin Chem Biol* **47**, 77-85, doi:10.1016/j.cbpa.2018.09.009
695 (2018).
696 59 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of
697 biomolecular interaction networks. *Genome research* **13**, 2498-2504,
698 doi:10.1101/gr.1239303 (2003).
699 60 Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a
700 New HHpred Server at its Core. *J Mol Biol* **S0022-2836**, 30587-30589 (2017).
701 61 Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist.
702 *Nucleic Acids Res* **36**, W465-469, doi:10.1093/nar/gkn180 (2008).
703 62 Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for
704 Timelines, Timetrees, and Divergence Times. *Molecular biology and evolution* **34**, 1812-
705 1819, doi:10.1093/molbev/msx116 (2017).

706